

By Jeanette W. Chung, Allison R. Dahlke, Cynthia Barnard, John O. DeLancey, Ryan P. Merkow, and Karl Y. Bilimoria

The Centers For Medicare And Medicaid Services Hospital Ratings: Pitfalls Of Grading On A Single Curve

DOI: 10.1377/hlthaff.2018.05345
 HEALTH AFFAIRS 38,
 NO. 9 (2019): 1523-1529
 ©2019 Project HOPE—
 The People-to-People Health
 Foundation, Inc.

ABSTRACT The star rating system for hospitals of the Centers for Medicare and Medicaid Services (CMS) pools all hospitals together and awards each institution one to five stars for quality, despite variation across hospitals in the numbers and types of measures they report. Thus, hospitals essentially are being evaluated differently, which affects the validity of quality comparisons. We considered the number and types of measures reported and the size of measure denominators to represent different forms of a “test,” and we used data from the December 2017 star ratings to show that hospitals took one of three general “test forms.” Hospitals taking the most extensive test form reported an average of forty-three measures, while those taking the least extensive test reported an average of twenty-two measures. These test forms were differentially associated with star ratings and hospital characteristics. Our results caution against pooling all hospitals together when assigning star ratings, and they demonstrate a feasible approach to segmenting hospitals into peer groups for evaluation by stakeholders such as CMS.

Jeanette W. Chung (jeanette-chung@northwestern.edu) is a research associate professor in the Surgical Outcomes and Quality Improvement Center (SOQIC), Department of Surgery, Feinberg School of Medicine, Northwestern University, in Chicago, Illinois.

Allison R. Dahlke is assistant director of population sciences at the University of Wisconsin Carbone Cancer Center, in Madison.

Cynthia Barnard is vice president of quality, Northwestern Memorial HealthCare, and an assistant professor in the Center for Healthcare Studies at the Feinberg School of Medicine, Northwestern University.

John O. DeLancey is a research fellow in urology, SOQIC, Department of Surgery, Feinberg School of Medicine, Northwestern University.

Ryan P. Merkow is faculty in SOQIC and an assistant professor in the Department of Surgery, Feinberg School of Medicine, Northwestern University.

Karl Y. Bilimoria is the J. B. Murphy Professor of Surgery in the Department of Surgery; the director of SOQIC, Department of Surgery, Feinberg School of Medicine, Northwestern University; and vice president of quality at Northwestern Memorial HealthCare.

The Overall Hospital Quality Star Ratings of the Centers for Medicare and Medicaid Services (CMS), introduced in 2016, have triggered concerns related to methodology and validity. There is concern about how individual quality indicators are weighted and combined into a composite measure and how stars are assigned to a hospital.^{1,2} There is also concern that hospitals constitute a heterogeneous group whose members might not be meaningfully compared to one another. Large, academic, tertiary care providers are grouped together with critical access and rural facilities in one rating pool, with great variability in the number and types of measures that hospitals report to CMS.

The star ratings are based on data submitted to CMS by hospitals that meet minimum data-reporting requirements. To receive full annual payment increases, hospitals must participate

in quality data reporting. Minimum reporting requirements vary across measure types, and not all measures may be appropriate for every hospital.³ A subset of fifty-seven measures was selected by CMS for use in development of the December 2017 Overall Hospital Quality Star Ratings. These measures represented seven measure groups: mortality, safety of care, readmission, patient experience, effectiveness of care, timeliness of care, and efficient use of medical imaging.⁴ To the extent that hospitals vary in the number of measures reported and in which measures they report, hospitals may metaphorically be characterized as taking different forms of a test.

If hospitals take different tests, then star ratings may be confounded by the test form. Higher ratings may reflect “easier tests” (for example, reporting fewer measures), while lower scores may reflect “harder tests” (reporting many mea-

sures). If star ratings are confounded by the test form, then patients, referring providers, payers, and other stakeholders who use the ratings in their decision making can be misled if they assume that higher hospital quality is correlated with a higher star rating.

Stakeholders, including the Medicare Payment Advisory Commission, have expressed concern that hospitals do not report a uniform set of measures to CMS.⁵ Moreover, the star ratings program pools providers from disparate populations and essentially grades them on a single curve, which is problematic if hospitals are taking different tests. In response to these problems, the Technical Expert Panel for the CMS star ratings program suggested grouping star ratings by hospital type or the number of measures or measure groups that hospitals report.⁶

If hospitals of the same type (for example, specialty, critical access, or rural hospitals) take similar tests, then grouping hospitals by type would make sense. However, reporting requirements are based not on hospital type, but rather on denominator volumes for each measure. A minimum number of relevant cases must be available to enable a hospital to report a quality measure, and hospitals of the same type may report different measures. Thus, the practical problem is this: Without mandating measure sets for reporting, how can profilers identify hospital peer groups based on the measures they report so that comparisons are made among hospitals that take the same (or similar) tests?

We addressed this question by exploring a simple approach that segments hospitals in the star ratings program into groups of hospitals that took similar tests—that is, reported similar measures, numbers of measures, and denominators. We identified these peer groups using cluster analysis of the measure denominators that hospitals reported. We then examined whether there was any systematic association between test forms and star ratings. Finally, we explored associations between test forms and hospital characteristics to gain insight into the nature of hospital peer groups based on test forms.

Study Data And Methods

DATA Data for these analyses came from the December 2017 CMS star ratings data set.⁷ Additional data came from the 2017 CMS Inpatient Prospective Payment System Payment Update Impact File⁸ and the 2016 American Hospital Association Annual Survey.⁹ In December 2017 hospitals received a star rating if they reported measures from at least three measure groups, of which at least one was an outcome group, and if all reported measure groups contained at least

three measures. Hospitals failing to meet these minimum public reporting requirements were excluded from receiving a star rating and from our analyses. A total of 3,692 hospitals received a star rating and were included in our analyses.⁴

IDENTIFYING DIFFERENT TEST FORMS We applied hierarchical cluster analysis to the CMS star ratings data to derive peer groups of hospitals based on similarity in the measures reported, the number of measures reported, and the denominator size for each measure. The fifty denominator variables included in our analysis are listed in online appendix A.¹⁰ For hospitals that did not report on a specific measure, we set the denominator for that measure to zero (that is, zero cases were reported). Denominators were standardized using z-scores before clustering. In our cluster analysis, hospitals were grouped together to minimize within-cluster variance in denominator sizes across all denominator measures. In this sense, clusters represent groups of hospitals that had similar patterns of measure reporting. We use the term “test forms” interchangeably with “cluster” to refer to these clusters of hospital peers taking the same test form. A description of our methods is in appendix B.¹⁰

For purposes of internal validation, we used one-way analysis-of-variance tests to assess the distinctiveness of the clusters of hospitals that had similar patterns of measure reporting, as identified by the cluster analysis. We tested whether the means of denominator sizes in each cluster were significantly different from one another for each of the fifty measures in the cluster analysis (shown in appendix C).¹⁰ Where Bartlett’s test for equality of variances indicated violation of the assumption of equal variances, we compared cluster means using pairwise *t*-tests for unequal variances, with Bonferroni correction for multiple comparisons.

We characterized each test form by the mean number of measures reported overall, mean number of measure groups reported, and mean number of measures within measure groups.

EVALUATING POSSIBLE CONFOUNDING OF STAR RATINGS BY TEST FORM We performed a chi-square test to assess whether test forms were systematically associated with star ratings. We also estimated a simple generalized ordered logistic regression model that regressed the stars awarded to a hospital on test form (cluster) to further characterize this association (see appendix E for additional methodological details).¹⁰

ASSOCIATIONS BETWEEN TEST FORMS AND HOSPITAL CHARACTERISTICS We used one-way analysis-of-variance tests and chi-square tests of association as described above to explore whether test forms were associated with partic-

ular types of hospitals (that is, whether patterns of measure reporting coincided with particular types of hospitals). The hospital characteristics we examined included average daily census; disproportionate share percentage; case-mix index; number of residents per ten beds; total bed size; total admissions volume; core-based statistical area type (rural, with a population of less than 10,000; micropolitan, with a population of 10,000–50,000; or metropolitan, with a population of more than 50,000), membership in the Council of Teaching Hospitals of the Association of American Medical Colleges, and Level I trauma center designation. In the case of continuous variables where Bartlett's test indicated unequal variances across groups, we again used pairwise *t*-tests with Bonferroni correction, as described above.

LIMITATIONS There were several limitations to our study. First, in cluster analysis, there are no objective criteria by which to choose one cluster solution over another.¹¹ We selected a three-cluster solution based on visual plots and indices that summarized how distinctive different numbers of clusters might be.

Second, while we demonstrated how cluster analysis can be used to identify groups of hospitals taking similar tests, our study intentionally stopped short of examining how grouping hospitals by test form would redistribute stars, as our objective was to assess differences in the tests taken.

Third, our cluster analysis included only the raw denominator data that CMS released in the December 2017 star ratings data set. If one were to implement this in policy practice, one might want to use averages of denominator data over multiple reporting periods.

Study Results

TEST FORMS IN THE DECEMBER 2017 STAR RATINGS DATA Cluster analysis suggested that the 3,692 hospitals that were accorded star ratings could be segmented into three clusters based on differences in overall patterns of measures reported and denominator volumes (for details, see appendix B).¹⁰ Internal validation results, reported in appendix C,¹⁰ show mean case volumes within each cluster for each of the fifty measure denominators. Bartlett's test was significant across all measures, which supported the distinctiveness of the clusters in representing different populations. Pairwise *t*-tests for unequal variances with Bonferroni correction for three pairwise comparisons per measure were all significant ($p < 0.001$), except for five outpatient measures. These findings, reported in appendix C,¹⁰ suggest that the three clusters extracted

from the analysis represent reasonably different test forms.

The mean number of measures, mean number of measure groups, and mean number of measures within each group were generally greater in cluster 1 than in the other two clusters (exhibit 1). Cluster 3 reported the lowest mean numbers of measures, measure groups, and measures within groups. Cluster 1 reported roughly twice as many measures as cluster 3 did, overall and for mortality, readmissions, effectiveness, and imaging measures. Among cluster 1 hospitals, the mean number of safety measures reported was 6.6, compared to 1.6 among cluster 3 hospitals.

THE ASSOCIATION BETWEEN TEST FORMS AND STAR RATINGS We found an overall association between test forms and star ratings (chi-square test of association: $p < 0.001$): Cluster 1, which reported the most measures and measure groups, accounted for 36 percent of one-star hospitals, 27 percent of two-star hospitals, 18 percent of both three- and four-star hospitals, and 33 percent of five-star hospitals (exhibit 2). Cluster 3, which reported the fewest measures and measure groups, accounted for 10 percent of one-star hospitals, 20 percent of two-star hospitals, 33 percent of three-star hospitals, 36 percent of four-star hospitals, and 22 percent of five-star hospitals.

Compared to cluster 2, cluster 1 was more likely to receive both the lowest ratings and the top rating (exhibit 3; see appendix E for estimates with standard errors and confidence intervals).¹⁰

EXHIBIT 1

Measures reported for the December 2017 Overall Hospital Quality Star Ratings of the Centers for Medicare and Medicaid Services (CMS), by test form cluster

	Mean number of measures reported			
	Cluster 1 (n = 826)	Cluster 2 (n = 1,811)	Cluster 3 (n = 1,055)	All (N = 3,692)
All measures	43.3	36.3	22.3	33.8
Measure groups	7.0	6.9	6.1	6.7
Mortality measures	6.8	5.0	3.0	4.9
Safety measures	6.6	4.7	1.6	4.2
Readmission measures	8.7	7.2	4.9	6.9
Effectiveness measures	10.2	8.9	5.6	8.3
Timeliness measures	5.4	5.9	4.5	5.4
Imaging measures	4.6	3.8	2.3	3.5

SOURCE Authors' analysis of December 2017 Overall Hospital Quality Star Ratings from CMS. **NOTES** The analysis included all 3,692 hospitals that received ratings. Cluster 1 consists of the hospitals that took the most extensive test form, while hospitals in cluster 2 took the moderate test form and those in cluster 3 took the least extensive test form. Bartlett's test for equality of variances was significant ($p < 0.001$). Thus, pairwise comparisons of means were conducted using *t*-tests for independent samples with unequal means. We corrected *p* values using Bonferroni's method for three pairwise comparisons per row. All pairwise comparisons were significant (Bonferroni-corrected $p < 0.001$) and are reported in appendix D (see note 10 in text). Although the text mentions a seventh measure group, patient experience, the CMS star ratings data set reports only a single denominator for all measures in this group. Therefore, it is omitted from the exhibit because of lack of variation.

EXHIBIT 2

Hospitals within each test form cluster, by star rating in 2017

Test form cluster	1 star (n = 261)	2 stars (n = 752)	3 stars (n = 1,189)	4 stars (n = 1,153)	5 stars (n = 337)	All (N = 3,692)
Cluster 1	36.4%	27.0%	18.1%	17.6%	32.6%	22.4%
Cluster 2	53.6	53.2	48.7	46.8	45.4	49.0
Cluster 3	10.0	19.8	33.2	35.7	22.0	28.6

SOURCE Authors' analysis of December 2017 Overall Hospital Quality Star Ratings from the Centers for Medicare and Medicaid Services. **NOTES** The clusters are explained in the notes to exhibit 1. Chi-square tests of association between all clusters and star ratings were significant ($p < 0.0001$).

Cluster 1 hospitals were 36 percent less likely than cluster 2 hospitals to receive 2–5 stars than 1 star (odds ratio: 0.64) and 25 percent less likely to receive 3–5 stars versus 1 or 2 stars (OR: 0.75). Cluster 1 hospitals were as likely as cluster 2 hospitals to receive 4 or 5 stars versus 1–3 stars (OR: 0.99). However, cluster 1 hospitals were 1.66 times more likely than cluster 2 hospitals to receive 5 stars versus 1–4 stars (OR: 1.66).

By contrast, cluster 3 hospitals were more likely than cluster 2 hospitals to receive higher ratings: Cluster 3 hospitals were more than three times as likely as cluster 2 hospitals to receive more than 1 star (OR: 3.32), more than twice as likely to receive 3–5 stars versus 1 or 2 stars (OR: 2.14), and 1.38 times more likely to receive 4 or 5 stars versus 1–3 stars (OR: 1.38). Cluster 3 hospitals were as likely as cluster 2 hospitals to receive the top rating (OR: 0.82).

ASSOCIATION BETWEEN HOSPITAL CHARACTERISTICS AND TEST FORMS To ascertain whether certain types of hospitals exhibit different patterns of measure reporting to CMS, we examined the association between various hospital characteristics and test forms. Cluster 1 hospitals, which reported the most measures and measure groups, were the largest and busiest hospitals: They had the highest average daily census (269.3), case-mix index (1.8), number of resi-

dents per ten beds (1.4), number of beds (449.1), and annual number of admissions (22,035.6) (exhibit 4). By contrast, cluster 3 hospitals, which reported the smallest number of measures and measure groups, had the lowest average daily census (25.6), case-mix index (1.3), number of residents per ten beds (0.3), number of beds (57.7), and number of annual admissions (1,635.3).

While 96 percent of cluster 1 hospitals were located in metropolitan core-based statistical areas, only 37 percent of cluster 3 hospitals were. Twenty-two percent of cluster 3 hospitals were located in micropolitan areas, and 42 percent were in rural areas. By contrast, only 4 percent of cluster 1 hospitals were located in micropolitan areas, and fewer than 1 percent were in rural areas. Twenty-two percent of cluster 1 hospitals were members of the Council of Teaching Hospitals, compared to 2 percent of hospitals in cluster 2. Only one hospital in cluster 3 was a member of the council. Twenty percent of cluster 1 hospitals were designated Level I trauma centers, compared to 3 percent and 1 percent of clusters 2 and 3 hospitals, respectively.

Discussion

The landscape of hospital quality reporting in the US is crowded with national, state, and local programs, as well as specialty-society and commercial programs. There is little alignment across programs, and evaluations of the same hospital across different programs are often inconsistent—involving different measures, data sources, and weighting schemes—which raises questions regarding the validity of evaluations.^{12,13} Previous researchers have found that hospitals awarded top ratings by the CMS star ratings program tended to be small rural hospitals, while large academic medical centers were overrepresented among the lowest-rated hospitals—which suggests that uncontrolled confounding may bias evaluations.^{2,14} Others have expressed concerns about data consistency,

EXHIBIT 3

Association between test form used and number of stars awarded in the hospital quality ratings of the Centers for Medicare and Medicaid Services (CMS) in December 2017

Star rating	Cluster 1	Cluster 3
2, 3, 4, or 5 stars versus 1 star	0.64***	3.32***
3, 4, or 5 stars versus 1 or 2 stars	0.75***	2.14***
4 or 5 stars versus 1, 2, or 3 stars	0.99	1.38***
5 stars versus 1, 2, 3, or 4 stars	1.66****	0.82

SOURCE Authors' analysis of December 2017 Overall Hospital Quality Star Ratings from CMS. **NOTES** The exhibit shows estimated odds ratios from a generalized ordered logistic regression model that regressed CMS stars on test form cluster, with cluster 2 as the reference (the clusters are explained in the notes to exhibit 1, and sample sizes are in exhibit 2). No additional covariates were included in the model. *** $p < 0.01$ **** $p < 0.001$

measure reliability and validity,¹² and lack of internal consistency across indicators of the same construct.¹⁵

Grouping similar hospitals together based on hospital type or reported measures has been proposed to enhance the validity of the CMS star ratings report.⁶ However, little attention has been focused on the practical problem of how to group hospitals. Conceptually, peer grouping should group together hospitals that take similar tests—that is, report similar measures and similar numbers of measures. Although there are associations between hospital types and patterns of measure reporting,² there are no program requirements for hospital types to report uniform sets of measures. Therefore, there is no way to ensure that hospitals of a particular type are all taking the same test.

In this study we demonstrated that hospitals can be grouped by the number and types of measures they report. We used hierarchical cluster analysis on denominator counts in the star ratings data to demonstrate the feasibility of segmenting hospitals into groups that have similar patterns of measure reporting. We found that there may be at least three patterns of measure reporting (what we have descriptively called test forms) in the December 2017 CMS star ratings data. We found that test forms were predictive of stars awarded and were associated with hospital characteristics. Hospitals that took the most extensive test form were more likely than those that took a moderate test form to receive the lowest as well as the highest star ratings and tended to be larger, high-volume, urban, tertiary-care teaching hospitals. Hospitals that took the most abbreviated test form were more likely than those that took a moderate test form to receive higher star ratings and tended to be smaller, low-volume, rural, nonteaching facilities.

The current star ratings methodology pools hospitals that take very different tests into a single distribution. This may facilitate the identification of apparent outliers, but it might not accurately reflect relative quality. As a result, this measurement approach may mislead patients, providers, and payers about where to seek high-quality care, and it may mislead hospitals in identifying opportunities for improvement—which could contribute to waste. Given the investment that CMS has made in its star ratings methodology, it would seem to be worth exploring the benefit of grouping hospitals into peer groups, perhaps based on test forms, to improve the validity and usability of the star ratings.⁶

The cluster analytic approach that we demonstrated is agnostic to hospital attributes and performance, and it provides a completely empirically based means of segmenting hospitals in the

EXHIBIT 4

Hospital characteristics by test form cluster, 2017

	Cluster 1	Cluster 2	Cluster 3
Average daily census (mean)	269.3	76.6	25.6
Disproportionate share percentage (mean)	29.6	30.5	31.9
Case-mix index (mean)	1.8	1.5	1.3
Residents per 10 beds (mean)	1.4	0.5	0.3
Number of beds (mean)	449.1	155.2	57.7
Annual admissions (mean)	22,035.6	6,741.9	1,635.3
CBSA type (%)			
Metropolitan	96.0	68.3	36.7
Micropolitan	3.8	22.9	21.5
Rural	0.3	8.8	41.7
COTH member (%)			
Yes	21.9	2.3	0.1
No	78.1	97.7	99.9
Level I trauma center designation (%)			
Yes	19.9	3.3	1.3
No	80.2	96.7	98.7

SOURCE Authors' analysis of the following data: December 2017 Overall Hospital Quality Star Ratings from the Centers for Medicare and Medicaid Services (CMS); CMS.gov. FY 2017 IPPS final rule and correction notice data files (note 8 in text); and 2016 American Hospital Association (AHA) Annual Survey (note 9 in text). **NOTES** The clusters and our correction of *p* values are explained in the notes to exhibit 1. The number of observations varies because of missing data in the CMS 2017 Inpatient Prospective Payment System (IPPS) Payment Update Impact File and the 2016 AHA Annual Survey. Bartlett's test for equality of variances was significant (*p* < 0.001). Thus, pairwise comparisons of means were conducted as explained in the notes to exhibit 1. All pairwise comparisons of means with Bonferroni correction were significant (*p* < 0.001) except for disproportionate share percentage (for which no pairwise comparisons were significant [*p* < 0.05]). Chi-square tests of association—used for core-based statistical area (CBSA) type, Council of Teaching Hospitals (COTH) membership, and trauma center designation—were all significant (*p* < 0.001). Appendix F shows standard deviations (see note 10 in text).

star ratings program, based on the similarity of the tests taken. The advantages of this approach are that peer groupings reflect actual similarity in test forms.

Our approach does have some drawbacks. The star ratings program changes over time with the addition or retirement of various measures or measure groups, and hospitals may report different measures over time. In response to these changes, the approach that we demonstrated could yield different numbers of hospital groups over time, and group memberships could change. While we regard this as appropriate, such changes could cause some difficulty for hospitals that seek stability in their peer group. It would also increase the complexity of the method for stakeholders and consumers in understanding how peer groups are determined and why they might change over time. Using a non-hierarchical clustering approach in which one specifies, *a priori*, how many clusters the data should be grouped into is one alternative that would enable clustering based on similarity in test form, while maintaining a constant number of peer groups from year to year.

In addition, our approach could make it more difficult to identify statistical outliers within hospital groups. However, the current approach of distribution-based norm-referenced grading could inhibit hospitals from setting and meeting rational, achievable improvement goals. When one is grading from a distribution, there will always be a fixed proportion of winners and losers, but the difference in performance between the two groups may be large or inconsequentially small from a practical standpoint. Moreover, relative performance is meaningless without reference to the mean and some measure of variability. If everyone performs poorly, five-star hospitals may still deliver poor-quality care, despite being the best of poor performers. If everyone excels, one-star hospitals may provide high-quality care, despite being the worst of a high-performing population. If a goal of performance measurement and reporting is to provide incentives for hospitals to reach a desired level of quality, this goal should be attainable. However, norm-referenced grading almost guarantees that this goal will never be met, because there will always be one- and five-star hospitals.

An alternative is criterion-referenced grading, in which hospitals would be judged by whether they met predefined performance targets. CMS could convene groups of stakeholders or measure developers to provide evidence or consensus-based performance targets for measures in Hospital Compare and other measurement and reporting programs. Criterion-referenced grading would enable hospitals to set performance goals that are not constantly moving with group improvement. As explicit goals, they provide clear targets to aim for and may help prevent flat-of-the-curve spending on quality improvement. Under criterion-based grading, a day when all hospitals meet or exceed performance targets on some metric is possible, and from a health system perspective, the proportion of laggards remaining becomes a meaningful measure

CMS may have an opportunity for meaningful improvement of the quality rating system by evaluating hospitals within well-defined peer groups.

of progress. Criterion-referenced measurement also facilitates self-improvement, because the target does not move with peer performance.

Conclusion

The hospitals that were included in the December 2017 CMS star ratings data appeared to be a heterogeneous group with respect to the measures, number of measures, and measure denominators they reported. Our analyses suggest that these groups of hospitals may be distinct populations, and we caution against pooling them in the assignment of stars. In addition, our findings suggest that CMS may have an opportunity for meaningful improvement of the quality rating system by evaluating hospitals within well-defined peer groups aligned on scope and type of relevant measures. Furthermore, criterion- instead of norm-based performance targets within these peer groups could be considered to accelerate and focus national quality improvement. ■

Karl Bilmoria is a member of the Centers for Medicare and Medicaid Services Hospital Quality Star Ratings on Hospital Compare Technical Expert Panel.

NOTES

1 Bilimoria KY, Barnard C. The new CMS hospital quality star ratings: the stars are not aligned. *JAMA*. 2016;316(17):1761–2.

2 DeLancey JO, Softcheck J, Chung JW, Barnard C, Dahlke AR, Bilimoria KY. Associations between hospital characteristics, measure reporting, and the Centers for Medicare & Medicaid Services Overall Hospital Quality Star Ratings. *JAMA*. 2017; 317(19):2015–7.

3 QualityNet. Archived participation resources, Hospital Inpatient Quality Reporting (IQR) Program, program guides: reference checklist for FY 2017 [Internet]. Baltimore (MD): QualityNet; 2016 Jan [cited 2019 Jul 22]. Available for download from: <https://www.qualitynet.org/dcs/ContentServer?c=Page&pagename=QnetPublic%2FPage%2FQnetTier4&cid=1144440979338>

4 QualityNet. Methodology, overall hospital ratings, methodology resources: comprehensive methodology report (v3.0) [Internet]. Baltimore (MD): QualityNet; 2017 Dec [cited 2019 Jul 22]. Available for download from: <https://www.qualitynet.org/dcs/ContentServer?c=Page&pagename=QnetPublic%2FPage%2FQnetTier3&cid=1228775957165>

5 Medicare Payment Advisory Commission. RE: File code CMS-4168-P [Internet]. Washington (DC): MedPAC; 2016 Sep 22 [cited 2019 Jul1]. Available from: http://medpac.gov/docs/default-source/comment-letters/20160922_medpac_pace_comment_sec.pdf?sfvrsn=0

6 Yale New Haven Health Services Corporation, Center for Outcomes Research and Evaluation. Hospital quality star rating on Hospital Compare: public input period: enhancements of the Overall Hospital Quality Star Rating [Internet]. Baltimore (MD): Centers for Medicare and Medicaid Services; 2017 Aug [cited 2019 Jul 2]. Available for download from: <https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/MMS/Downloads/Hospital-Compare-PC-2017.pdf>

7 QualityNet. Statistical Analysis System (SAS) Package, overall hospital ratings, December 2017 SAS package resources [Internet]. Baltimore (MD): QualityNet; [cited 2019 Jul 22]. Available for download from: <https://www.qualitynet.org/dcs/ContentServer?c=Page&pagename=QnetPublic%2FPage%2FQnetTier3&cid=1228775958130>

8 CMS.gov. FY 2017 IPPS final rule and correction notice data files [Internet]. Baltimore (MD): Centers for Medicare and Medicaid Services; [last modified 2017 Mar 7; cited 2019 Jul 1]. Available for download from: <https://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/>

9 American Hospital Association. Annual survey database. Chicago (IL): AHA; 2016.

10 To access the appendix, click on the Details tab of the article online.

11 Blashfield RK, Aldenderfer MS. The methods and problems of cluster analysis. In: Nesselroade JR, Cattell RB, editors. *Handbook of multivariate experimental psychology*. 2nd ed. New York (NY): Plenum Press; 1988. p. 447–73.

12 Shahian DM, Mort EA, Pronovost PJ. The quality measurement crisis: an urgent need for methodological standards and transparency. *Jt Comm J Qual Patient Saf*. 2016; 42(10):435–8.

13 Hwang W, Derk J, LaClair M, Paz H. Finding order in chaos: a review of hospital ratings. *Am J Med Qual*. 2016;31(2):147–55.

14 Wang DE, Tsugawa Y, Figueroa JF, Jha AK. Association between the Centers for Medicare and Medicaid Services hospital star rating and patient outcomes. *JAMA Intern Med*. 2016;176(6):848–50.

15 Hu J, Jordan J, Rubinfeld I, Schreiber M, Waterman B, Nerenz D. Correlations among hospital quality measures: what “Hospital Compare” data tell us. *Am J Med Qual*. 2017; 32(6):605–10.